Title: Comparability and Reliability Considerations of Adequate Yearly Progress

Authors and Affiliations: Kimberly S. Maier, Tapabrata Maiti, Sarat C. Dass, and Chae Young Lim, Michigan State University

Background: The No Child Left Behind (NCLB) Act of 2001 leaves much of the details of establishing and implementing an accountability system up to the individual states, with the requirement for the system to be valid and reliable (NCLB, 2002). Determination of how well schools and districts are doing is based largely on the classification of student performance on standardized assessments. However, making classification decisions about students has been part and parcel of research on testing long before NCLB was implemented. Some early work on error in such classification systems considered the problem with a loss function framework (Traub & Rowley, 1980). Other work included focused studies of misclassification using alternate test forms (Livingston & Lewis, 1995) or the application of generalizability theory (Brennan & Kane, 1977). Large sample sizes were demonstrated to produce more accurate and precise classification (Yen, 1997).

The reliability of a classification of student performance necessarily impacts AYP determinations for groups, be it subgroups within schools, schools themselves, or districts. In implementing NCLB, individual states considered the error associated with an aggregated decision and the importance of group size. In general, state-specific treatment of classification error proceeded along one of two routes. About 80% of states use conventional confidence intervals (CI) to express uncertainty, while some states use the standard error of measurement (SEM; U.S. Department of Education, 2010). The estimate of AYP and these approaches to treat misclassification will be compared to an alternate estimator that will be described below.

Purpose: The purpose of this study is to develop an estimate of Adequate Yearly Progress (AYP) that will allow for reliable and valid comparisons among student subgroups, schools, and districts. A shrinkage-type estimator of AYP using the Bayesian framework is described. Using simulated data, the performance of the Bayes estimator will be compared to currently-used non-Bayes estimator. While it is likely that NCLB will either experience a major overall or go away altogether, decisions will still have to be made about children's academic performance, and these systems will likely continue to use assessment performance as an indicator of academic performance. The estimator developed here has application to such accountability measure.

Significance: Differences in group sample sizes pose a difficult challenge to reliable estimates of AYP. For example, Table 1 shows the distribution of school district sizes in the state of Michigan. Figure 1 shows an example of how variable public school students' scores are in districts across the state. Highly variable district sizes make the direct estimators heterogeneous, which in term compromises their value for drawing conclusions and making inferences. Variability such as that shown here can be properly reflected when a so called "shrinkage" estimator is used. This estimator is an alternative to the commonly used "direct" estimator (the estimate is based on information specific only to the district or subgroup considered). A shrinkage-type approach has been used in hierarchical models (e.g., Raudenbush & Bryk, 2001) to estimate the relationship between variables while taking account of the different levels of observed data (e.g., students, classrooms, schools). A shrinkage-based technique for AYP determination would allow for comparisons of AYP across schools and districts.

(Please insert Table 1 here)

(Please insert Figure 1 here)

Statistical Model: Below, a Bayes estimator of AYP conformance is described and compared to a currently used estimator. Prior to the development of these estimators, the general procedure of AYP determination is described. Given the latitude NCLB gave to states about the procedure for determining AYP, it's important to note that the procedure described below is not that used by all states, but is merely meant to serve as a representative example. The procedures implemented by the state of Michigan are occasionally used to help illustrate a point. Regardless of the specific procedure used by a state, the estimators described below are applicable.

For each i =1, 2, $\cdots$,n, let school i comprise m_i students and a proportion p_i of students who are proficient; a student is declared proficient if his/her score on a particular test is equal to or above the State specific target. Thus, $100p_i$ is the percent proficient (see Michigan Department of Education, 2011) for the *i-th* school. This proportion p_i is unknown and is estimated for school i using $\hat{p}_i$, the observed proportion of students meeting the AYP criteria. Since a student's score on a test is typically a random variable, it follows that $\hat{p}_i$ has measurement error built in depending on all such students' scores, and thus, is also a random quantity. AYP classification is a high-stakes decision and it is important to estimate each p_i as reliably as possible by taking into account all possible sources of uncertainty.

More specifically, let $\hat{p}_{ik}$ be the observed proportion of proficient students in the *k-th* subgroup in school i, for k =1, 2, ..., K corresponding to the grade levels in school i. Assume that the number tested is m_{ik}, the true proportion of proficient students is p_{ik}, and the target is p_k^0 for the *k-th* subgroup. The AYP score for school i is determined as:

$$AYP\ score = 100\sum_{k=1}^{K}\frac{m_{ik}}{m_i}\left(\hat{p}_{ik} - p_k^0\right), \tag{1}$$

which is then compared to the threshold value of 0. School I is declared to have met the State objective if the weighted score is equal to or above 0. More generally, the above procedure is applied separately to the different subjects (i.e., reading, mathematics) and the subject-specific AYP score is compared to 0. Most states use a combination rule to determine whether the school meets AYP (for an example of such a rule for Michigan, see Michigan Department of Education, 2011, for more details).

Referencing (1), the key statistic used to determine the final AYP score is the proportion of proficient students $\hat{p}_{ik}$ in each subgroup. Also, the mean AYP score is obtained by replacing the statistics $\hat{p}_{ik}$ by its unknown true value p_{ik}. The school is truly proficient if the unknown mean AYP score is greater than or equal to 0. Due to uncertainty in the $\hat{p}_{ik}$s, there can be two types of errors associated with the decision made: A school can be declared to have met AYP when the true mean score is below 0 or the school can be declared not to have met AYP when the true mean score is above 0. These two errors, called false-positive and false-negative, can be minimized if the true but unknown p_{ik} values are estimated accurately. In the subsequent discussion, we focus on one $p_{ik} \equiv p_i$, noting that the models elicited for one such p_i can be easily generalized to include multiple p_{ik}s. Furthermore, currently practiced procedures any states can be suitably adapted on a case by case basis under the proposed general framework.

To determine whether a school meets AYP or not, many states have adopted the confidence interval approach (Forte-Fast & Erpenbach, 2004; Marion et al., 2002; U.S. Department of Education, 2010). The confidence interval procedure is as follows: Based on the observed proportion $\hat{p}_i$ for school i, construct the standard error of the proportion, $\mathrm{SE}(\hat{p}_i)$, and say, the 95% upper confidence interval for p_i using the normal approximation. This upper one-sided confidence interval is given by $\left[0, \hat{p}_i + z_{0.95}SE\left(\hat{p}_i\right)\right]$, where $z_{0.95}$ is a z-value such that $\Phi(z_{0.95})$ = 0.95 for the cumulative distribution function of the standard normal distribution, Φ. This approximation is valid only when the student size m_i is large; thus, when m_i is small, either the estimates are not reported or if reported, the estimates are known to be unreliable for determining AYP conformance. Nevertheless, the above confidence interval is compared to the target level, *c*, which is determined by the state. School *i* meets AYP if its upper tail is above the target level, that is, if $\hat{p}_i + z_{0.95}SE\left(\hat{p}_i\right) \geq c$.

The use of the upper confidence limit, although easy to interpret and justify, presents some serious difficulties. When the sample size m_i is small, the standard error expression $SE\left(\hat{p}_i\right)$ is only a crude approximation to the true standard error. In fact, it is much larger than the true value resulting in a higher upper confidence limit. The normal approximation is also quite unreliable in this case, and thus, the use of $z_{0.95}$ as a critical value may be suspect. More subtly, if this upper confidence criterion is used to determine AYP decisions for $n > 1$ schools *simultaneously,* it will produce a significant upward bias. This will be true even for large m_i values. To demonstrate this last fact, we consider the scenario of estimating the proportion of schools in a district that meet AYP, $\theta = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\left(p_i \geq c\right)$where I is the indicator function which takes the value 1 if $p_i \geq c$, and 0, otherwise. Note that $0 < \theta < 1$.

The upper confidence level approach gives rise to an estimate of θ, $\hat{\theta}_U$, given by $\hat{\theta}_U = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\left(\hat{p}_i + z_{0.95}SE(\hat{p}_i) \geq c\right)$. The estimate $\hat{\theta}_U$ has an upward bias. This results from the introduction of the term $z_{0.95}SE(\hat{p}_i)$which includes some schools that have actual proportion p_i is less than *c*. In other words, $\hat{\theta}_U$ will overestimate the true proportion θ, and will favor certain schools even when they have not performed at the desired level. Second, if m_i is small, the confidence intervals could be wider due to large variance for small subgroups, thus accentuating the upward bias. In addition, $\hat{\theta}_U$ is subject to higher variability since $SE(\hat{p}_i)$ is again estimated (the true value is $SE(p_i)$). Further, $SE(p_i)$ itself could be large particularly for small m_i.

Alternative estimators of p_i and θ are motivated from the Bayesian perspective. Assume for the moment that m_i is large so that $\hat{p}_i$ is approximately normally distributed with mean p_i and variance $\sigma_i^2 \equiv p_i(1-p_i)/m_i$ (note that the model below can be extended to encompass the case of small m_i and other flexible distributions for p_i, but these extensions are outside of the scope of the present study). We consider the hierarchical (multilevel) model given by

$$\hat{p}_i \overset{ind}{\sim} N(p_i, \sigma_i^2), \text{ and} \tag{2}$$

$$\mathrm{logit}(p_i) \overset{iid}{\sim} N(\mu, \tau^2), \tag{3}$$

for $i = 1, 2, \ldots, n$; in (2), '*ind*' refers to independent and in (3), '*iid'* refers to independent and identically distributed. The logit transformation of p_i is $\mathrm{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$.

Using Bayes theorem, the posterior of each p_i is determined by the above hierarchical model specification and is given, up to a proportionality constant, by

$$\pi(p_i \mid \hat{p}_i, \mu, \tau) \propto \pi(\hat{p}_i \mid p_i)\ \pi(p_i, \tau) \tag{4}$$

where $\pi(\hat{p}_i \mid p_i) = (2\pi\sigma_i^2)^{-1/2} e^{-(p_i-\hat{p}_i)^2/(2\sigma_i^2)}$ from the first stage of (4) and

$\pi(p_i \mid \mu, \tau) = (\partial \mathrm{logit}(p_i)/\partial p_i)(2\pi\tau^2)^{-1/2} e^{-(p_i-\mu)^2/(2\tau^2)}$ from the second stage of (4). The alternative Bayes estimator of θ is

$$\hat{\theta}_B = \frac{1}{n}\sum_{i=1}^{n} P^*(p_i \geq c) \qquad (5)$$

Where the probability P^* is computed with respect to the posterior distribution of p_i, given $\hat{p}_i$ in (4). The utility of this estimator in comparison to the confidence interval estimator is demonstrated in the next section.

Research Design: A simulation procedure will be used to compare the performance of $\hat{\theta}_B$ and $\hat{\theta}_U$. The study will produce 500 replicates will be used to compute the probability P^* in (5) as well as the expectation and variance under P^* using Monte Carlo. The following factors will be varied in the study: number of students per school, number of schools per district, and AYP cut-off thresholds, *c*. Two different numbers of students in a school will be considered, the values of which will represent a small and a modest number of students. Two choices for the number of schools per district will also be considered. And finally, four different values of the true proportion of schools in the district meeting AYP will be considered.

Findings / Results: A simulation procedure was carried out to demonstrate the improvement of $\hat{\theta}_B$ over $\hat{\theta}_U$ based on the hierarchical model in (2) and (3). The number of students in a school were varied to be $m = 30$ and 200. The number of schools in a district were varied to be $n = 10$ and $n = 15$. The AYP cut-off c was set to 0.7 = 70% and μ values for the true proportion of schools in a district meeting AYP were, $\theta = \{0.5, 0.599, 0.705, 0.813\}$. The prior variance was taken to be τ^2 =1. The true and observed proportions, p_i and $\hat{p}_i$, were generated based on the above parameter specifications.

(Please insert Table 2 here)

Bias, variance, and mean squared error (MSE) were used as indices of performance for each of the estimators and the numerical results presented in Table 2. The results indicate the superiority of the Bayes estimator $\hat{\theta}_B$ compared to the direct (non-Bayes) estimator $\hat{\theta}_U$. The inclusion of the confidence interval width led to the increased bias of $\hat{\theta}_U$, contributing to a higher MSE. The *i-th* term of the estimate $\hat{\theta}_U$ is based on the maximum likelihood estimator $\hat{p}_i$ which uses the information only from the *i-th* school. Thus, $\hat{p}_i$ has high variability, particularly if m_i is small. On the other hand, the *i-th* term in the Bayes estimator $\hat{\theta}_B$ is derived from a combination of $\hat{p}_i$ and the overall mean μ. The overall mean contribution has the effect of reducing the variance of $\hat{\theta}_B$ and thereby increasing its stability.

Conclusions: The proposed Bayes estimator for AYP classification offers an attractive alternative to current methods, as measured by performance in terms of bias, MSE, and variance reduction.

Appendix A. References

Brennan, R. L. and Kane,M.T. (1977). An index of dependability of mastery tests. *Journal of Educational Measurement, 14*(3), 277-289.

Forte Fast, E. and Erpenbach, W. J. (2004). *Revisiting statewide educational accountability under NCLB: A summary of requests in 2003-2004 for amendments to state accountability plans*. Washington, D. C.: Council of Chief State Schools Officers.

Livingston, S.A. and Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179-197.

Marion, S. F., White, C., Carlson, Erpenbach, W. J. , Rabinowitz, S. and Sheinker, J. (2002). *Making valid and reliable decisions in the determination of adequate yearly progress: A paper in the series: Implementing the state accountability system requirements under the No Child Left Behind Act of 2001*. Washington D.C.: Council of Chief State Schools Officers.

Michigan Department of Education. (2011). *Higher expectations cause more schools to not make adequate yearly progress in 2011*. Retrieved from www.michigan.gov/mde.

No Child Left Behind Act of 2001, Pub. L. No. 107-110 section 115 Stat. 1425 (2002).

Raudenbush S.W. and Bryk, A.S. (2001). *Hierarchical Linear Models,* 2nd ed. Thousand Oaks, CA: Sage

Traub, R.E. and Rowley, G.L. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement, 4,* 517-545.

U.S. Department of Education (2010). *State and Local Implementation of the No Child Left Behind Act, Volume IV - Accountability Under NCLB: Final Report*. Washington D.C.: Author.

Yen, W.M. (1997). The technical quality of performance assessments: Standard errors of percents of pupils reaching standards. *Educational Measurement, 16*(3), 5-15.

Appendix B. Tables and Figures

Table 1: Number of public school 4th graders in school districts in the State of Michigan

Range of Numbers of Students	Number of School Districts
1 ~ 100	210
100 ~ 200	124
200 ~ 300	62
300 ~ 500	39
500 ~ 1000	30
1000 ~	11

Table 2: Simulation Results

$n = 10$ $\theta =$	0.500		0.599		0.705		0.813	
$m = 30$	$\hat{\theta}_U$	$\hat{\theta}_B$	$\hat{\theta}_U$	$\hat{\theta}_B$	$\hat{\theta}_U$	$\hat{\theta}_B$	$\hat{\theta}_U$	$\hat{\theta}_B$
Bias	0.217	-0.054	0.200	-0.048	0.167	-0.039	0.113	-0.027
Var	0.020	0.00009	0.016	0.00007	0.011	0.00004	0.007	0.00002
MSE	0.067	0.003	0.056	0.002	0.039	0.002	0.020	0.001
$m = 200$	$\hat{\theta}_U$	$\hat{\theta}_B$	$\hat{\theta}_U$	$\hat{\theta}_B$	$\hat{\theta}_U$	$\hat{\theta}_B$	$\hat{\theta}_U$	$\hat{\theta}_B$
Bias	0.092	-0.056	0.097	-0.050	0.086	-0.040	0.056	-0.028
Var	0.026	0.00008	0.020	0.00006	0.015	0.00003	0.011	0.00001
MSE	0.034	0.003	0.029	0.003	0.022	0.002	0.014	0.001

$n = 15$ $\theta =$	0.500		0.599		0.705		0.813	
$m = 30$	$\hat{\theta}_U$	$\hat{\theta}_B$	$\hat{\theta}_U$	$\hat{\theta}_B$	$\hat{\theta}_U$	$\hat{\theta}_B$	$\hat{\theta}_U$	$\hat{\theta}_B$
Bias	0.232	-0.054	0.201	-0.048	0.159	-0.039	0.116	-0.027
Var	0.012	0.00005	0.011	0.00004	0.007	0.00003	0.005	0.00001
MSE	0.066	0.003	0.051	0.002	0.033	0.002	0.018	0.001
$m = 200$	$\hat{\theta}_U$	$\hat{\theta}_B$	$\hat{\theta}_U$	$\hat{\theta}_B$	$\hat{\theta}_U$	$\hat{\theta}_B$	$\hat{\theta}_U$	$\hat{\theta}_B$
Bias	0.098	-0.056	0.101	-0.049	0.083	-0.040	0.061	-0.028
Var	0.015	0.00005	0.014	0.00004	0.010	0.00002	0.008	0.00001
MSE	0.025	0.003	0.024	0.002	0.017	0.002	0.011	0.001

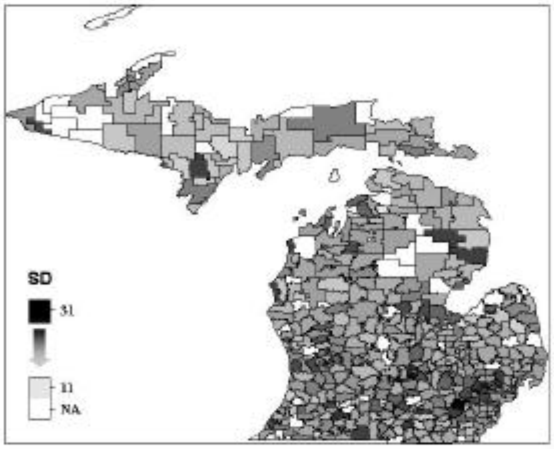

Figure 1: Standard deviation of public school students' assessment scores in available school districts in the State of Michigan.